Blame the wealthy? Examining the possible causal link between economic output and CO2 emissions

# 1. Introduction to the source data

I have downloaded the indicators for the analysis from the World Bank’s World Development Indicators website[[1]](#footnote-1). I have gathered the following variables:

* my outcome variable is *Carbon dioxide (CO2) emissions (total) excluding LULUCF[[2]](#footnote-2) (Mt CO2e)*, which will be later transformed to a tons per capita measure (CO2 for short);
* my causal variable is *GDP per capita, PPP (constant 2021 international $)* (GDP p.c. for short);
* I have identified 4 possible confounder variables (with indication of the possible confounding mechanism):
  + energy intensity of the economy (*Energy use (kg of oil equivalent) per $1,000 GDP (constant 2021 PPP),* energy intensity for short): as an economy gets more energy intense, it might have both a higher GDP and higher CO2 emissions;
  + sectoral composition of the economy (*Industry (including construction), value added (% of GDP)*, industry share for short): industrialization of a country might entail higher GDP and higher emissions;
  + renewable energy share in energy output (*Renewable electricity output (% of total electricity output)*, renewable share for short): higher use renewable sources might entail lower emissions, but higher GDP (as more developed economies turn towards renewables);
  + urbanization (*Urban population (% of total population)*,): as more and more people move to the cities, emissions may get lower (as people need to commute much less), but GDP may grow (as people perform higher value-added jobs in large cities);
* I have also downloaded the *Population, total* variable to calculate per capita CO2 emissions (population for short).

As for the outcome, note that it only measures CO2 emissions, rather than total greenhouse gas (GHG) emissions, so it is an imperfect measure of GHG emissions. I have chosen this variable rather than total GHG emissions in CO2 equivalent units, as the task description explicitly asked for this. After transforming to a tidy long format, my raw data had 6944 country-year observations (217 countries observed through 32 years, from 1992 until 2023). The percentage of non-missing observations per variable are presented in Table 1.

**Table 1: Percentage of non-missing country-year observations per variable**

|  |  |
| --- | --- |
| Variable name | Non-missing percentage |
| CO2 | 93.55 |
| GDP p.c. | 89.52 |
| Energy intensity | 45.97 |
| Industry share | 85.27 |
| Renewable share | 75.06 |
| Urban population share | 99.08 |
| Population | 100.00 |

# 2. Data cleaning

As I could do nothing with observations where either the outcome or the causal variables were missing, I dropped all such countries. This meant dropping 37[[3]](#footnote-3) out of the 217 countries[[4]](#footnote-4), leaving 180 countries in my dataset (meaning 5770 country-year observations).

Next, I looked at my possible confounders. As both energy intensity and renewable share still had a large proportion of missing values, I decided not to work with them further. Unfortunately, the remaining 7.6% missing values in industry share affected roughly a hundred of the countries still in my sample, so instead I decided to go on with the urban population share, which had no missing values at this point. After having filtered my sample, I added the CO2 per capita variable by simply dividing CO2 by population.

# 3. Descriptive statistics

Table 2 presents the summary statistics on my three main variables. We can see already that both GDP p.c. and CO2 p.c. have very skewed[[5]](#footnote-5) distributions with extreme values. This indicates that it is indeed reasonable to log-transform these variables.

**Table 2: Summary statistics of the untransformed variables**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | SD | Min. | 5% | 10% | 25% | 50% | 75% | 90% | 95% | Max. |
| GDP p.c. | 2.13e4 | 2.40e4 | 510.82 | 1492.49 | 2172.29 | 4160.62 | 1.24e4 | 3.11e4 | 5.43e4 | 6.65e4 | 1.74e5 |
| CO2 p.c. | 5.12 | 9.90 | 0.00 | 0.06 | 0.13 | 0.56 | 2.30 | 6.54 | 11.36 | 18.13 | 202.87 |
| Urban pop. % | 56.17 | 23.44 | 6.29 | 18.38 | 23.59 | 36.53 | 56.21 | 74.64 | 84.42 | 94.16 | 100.00 |

1. World Bank (2025). *World Development Indicators*. <https://databank.worldbank.org/source/world-development-indicators> [↑](#footnote-ref-1)
2. LULUCF means land use, land-use change and forestry. I have not found such a variable that would include LULUCF in CO2 emissions, thus I have gone with excluding this. [↑](#footnote-ref-2)
3. The following countries were dropped (by country code): AFG, AND, ASM, BTN, CHI, CUB, CUW, CYM, DJI, ERI, FRO, GIB, GRL, GUM, IMN, LBN, LIE, MAF, MCO, MNE, MNP, NCL, PRK, PSE, PYF, SMR, SRB, SSD, SXM, SYR, TCA, TON, VEN, VGB, VIR, XKX, YEM. [↑](#footnote-ref-3)
4. Note that I have downloaded the data from the website with selecting all units that were classified as countries. This meant data for some autonomous territories as well – thus the larger number than the official 195. [↑](#footnote-ref-4)
5. This is also shown by the histograms per year, omitted from this report because of spatial constraints. Please find the figures in the Jupyter notebook. [↑](#footnote-ref-5)